A linear response theory based method for prediction of large scale protein conformational changes upon ligand binding

Rajat Punia¹ and Gaurav Goel^{1,*}

¹Department of Chemical Engineering, Indian Institute of Technology Delhi, New Delhi, India

*goelg@chemical.iitd.ac.in

ABSTRACT

Prediction of ligand-induced protein conformational transitions is a challenging task due to a large and rugged conformational space, and limited knowledge of probable direction(s) of structure change. These transitions can involve a large scale, global (at the level of entire protein molecule) structural change and occur on a timescale of milliseconds to seconds, rendering application of conventional molecular dynamics simulations prohibitive even for small proteins. We have developed a computational protocol to efficiently and accurately predict these ligand-induced structure transitions solely from the knowledge of protein apo structure and ligand binding site. Our method involves a series of small scale conformational change steps, where at each step linear response theory is used to predict the direction of small scale global response to ligand binding in the protein conformational space (d_{LRT}) followed by construction of a linear combination of slow (low frequency) normal modes (calculated for the structure from the previous step) that best overlaps with d_{LRT} . Protein structure is evolved along this direction using molecular dynamics with excited normal modes (MDeNM) wherein excitation energy along each normal mode is determined by excitation temperature, mode frequency, and its overlap with d_{LRT} . We show that excitation temperature (ΔT) is a very important parameter that allows limiting the extent of structural change in any one step and develop a protocol for automated determination of its optimal value at each step. We have tested our protocol for three protein–ligand systems, namely, adenylate Kinase – di(adenosine-5')pentaphosphate, ribose binding protein – β -D-ribose, and DNA β -glucosyltransferase – uridine-5'-diphosphate, that incorporate important differences in type and range of structural changes upon ligand binding. We obtain very accurate prediction for not only the structure of final protein–ligand complex (holo-structure) having a large scale conformational change, but also for biologically relevant intermediates between the apo and the holo structures. Moreover, most relevant set of normal modes for conformational change at each step are an output from our method, which can be used as collective variables for determination of free energy barriers and transition timescales along the identified pathway.

Introduction

Conformational transitions are central to the biological functioning of proteins.^{1,2} Recent studies have suggested that such transitions are encoded in the protein fold and are triggered by external stimuli.^{3,4} In particular, ligand-induced conformational transitions play an important role in several biological pathways, viz. signalling,⁵ membrane transport,⁶ protein synthesis,⁷ enzyme catalysis,⁸ active transport by motor proteins,⁹ and allosteric regulation,¹⁰ making knowledge of protein structural response upon ligand binding very important for firmly establishing structure-function relationships.¹¹ Further, prediction of ligand-induced protein structural transitions are very important for structure-based drug discovery approaches.¹² For example, accounting for receptor conformational flexibility has been shown to improve prediction of both the structure and the binding affinity of the protein-ligand complex.^{2,12,13}

Advancement in computer hardware and simulation algorithms coupled with significant progress in experimental techniques for structure determination has provided several insights into a dynamic nature of protein structure – function relationships spanning from local (at binding site) to global (at the level of entire protein molecule to protein assemblies). However, prediction of large-scale conformational transitions of proteins including identification of functionally relevant intermediates (on-pathway), either experimentally or computationally, still presents a formidable task (for example, see recent review articles on the subject^{11,12}). These structural transitions span at least two orders of lengthscale (from atomic, ~ 1 Å, to entire protein molecule, ~ 100 Å) and typically involve large barriers on a rugged free energy landscape (FEL), leading to transition timescales of milliseconds to seconds and conformations trapped in a local minima. All atomistic molecular dynamics (MD) simulations based on the system Hamiltonian, although ideally suited to probe details on the required atomic lengthscale, are made intractable by large range of timescales involved. Several methodologies proposed to overcome this timescale problem, such as reaction path methods and potential smoothing approaches, have been shown to generate atomically detailed functional trajectories of protein transitions that provide information on structure of biologically relevant intermediates and end states as well as free energy barriers and transition timescales.¹⁴ These techniques, referred to as low-throughput MD methods, are still computationally intensive and this resource requirement typically increases exponentially with sum total of relevant degrees of freedom for protein dynamics.^{15,16}

An alternate approach involves the use of coarsegrained dynamics in the subspace of protein degrees of freedom, with the normal modes analysis (NMA) being the most prominent. Numerous studies have established that protein structural transitions upon ligand binding can be described by a small set of slow normal modes of protein fluctuations in its unbound (apo) state.^{3,4,17–20} Emergence of highly simplified elastic network models (ENMs) to obtain these normal modes²¹ from

a single structure, for example a crystallography structure, led to widespread use of NMA in flexible docking, targeted conformational search, and prediction of functional motions of proteins. Several enhanced sampling techniques that accelerate protein dynamics along these slow modes combined with use of coarsegrained experimental information on the structure of protein-ligand complex as constraints have shown success in predicting the holo (ligand-bound) state of protein from its apo (ligand-unbound) state.²² For example, structural displacement along a linear combination of low-frequency modes of the apo state, where coefficients of linear combination were determined using a small set of distance constraints of the end state, was used to predict the structure of the end state.²³ Perahia's group developed an enhanced sampling method, MDeNM (molecular dynamics with excited normal modes), in which protein is kinetically excited along the selected subset of modes.²⁴ Using this method, end conformations are obtained with the prior knowledge of dominant normal modes involved in conformational transitions upon ligand binding. Since overall protein structure change can be non-linear, whereas only linear structure change is possible by displacement along a given set of normal modes, techniques based on small displacement driven along the predicted direction and a re-calculation of low frequency motion from the obtained structure are expected to be better suited for determination of large-scale structural changes. Tama et al. used the linear combination of slow modes to iteratively deform the structure towards target guided by low resolution electron density map of target structure.²⁵ Wu and co-workers has accurately predicted the folded state of GB1 starting from its partially folded state using cryo-electron microscopy map restrained self-guided Langevin dynamics.^{26,27} All the above approaches use NMA based enhanced sampling methods guided by either low resolution structural information on end state or prior knowledge of transition pathway.

Oftentimes prior knowledge on end state or reaction coordinate is not available, for example in virtual drug screening, where incorporation of receptor flexibility notably improves screening accuracy.¹³ One option involves displacement along a predefined number of lowest frequency normal modes combined by energy minimization of the docked ligand-protein complex to predict holo state solely from its apo state.^{28,29} Computational cost of these methods increases geometrically (2^x) with the number of modes (x) selected for optimization. Moreover, inclusion of irrelevant modes would add noise in the energy calculations performed to identify the most stable ligand-bound state³⁰. Therefore, to achieve high accuracy and computational efficiency, the most important choice in protein holo structure prediction methods is the selection of the relevant degrees of freedom¹². Cavasotto and co-workers introduced a heuristic measure to determine relevance of normal modes, and thus enable exclusion of irrelevant modes³⁰. This involved generation of multiple receptor conformers by perturbing structure along random linear combinations of selected relevant modes, followed by docking

and energy optimization to identify the most stable ligand-bound state³⁰. Another option is to identify the specific set of normal modes from the knowledge of intrinsic dynamics of protein in free state and perturbation caused by ligand binding^{3,31}. Few recent studies attempted to model this predisposition of conformational transitions without any knowledge on end conformation. Zheng and Doniach predicted the conformational change in terms of Hessian matrix of protein in its unbound state and ligand induced structure distortions in binding pocket⁹. Ikeguchi and co-workers have shown that classical linear response theory (LRT) can predict the coupling relation between ligand-induced force on protein and protein conformational response³². Notably, the governing equations derived by Zheng and Doniach for predicting direction of conformational change is quite similar to the LRT. The direction of conformational change predicted showed a high correlation with observed direction of conformational change. However, prediction of final state using these equations is not straightforward because protein conformational change in general is non-linear and these equations do not accurately predict the extent of conformational change. Tamura and Hayashi have shown that sole application of LRT failed to accurately predict the conformational change in N-calmodulin which undergoes large scale non-linear conformational transitions upon ligand binding¹⁶. Moreover, they have shown that the direction of protein response changes along the transition path and LRT can be used in a sequential manner to identify the relevant coordinate of conformational change in each new structure identified along the transition path.

Apart from the prediction of protein holo state from its apo state, the information on apo $\rightarrow$ holo transition pathway and short-lived intermediate states is crucial for understanding the biological functioning mechanisms of proteins and developing rational intervention strategies³³. Accurate prediction of transition pathway requires a converged sampling of free energy landscape along the CVs describing the transition path which is non-trivial and computationally expensive. However, several coarse-grained morphing methods have been proposed for generation of transition pathways connecting apo and holo states. MolMovDB³⁴, which performs linear interpolation between apo and holo state in Cartesian space; NOMAD-Ref³⁵, which uses ENM-based normal modes to interpolate inter-residue distances between specified end states; MinAction-Path³⁶, which solves action-minimization problem to find the crossing point of trajectory solutions generated from the two end states; NMSim³⁷, which performs biased backbone motion in slow-modes subspace followed by favourable rotamer state search for side-chains; and iMODS³⁸, which performs geometric interpolation on dihedral angle space, are some of the coarse-grained methods available for use via web-servers. Recently, Orellana and co-workers have shown the application of hybrid elastic-network Brownian dynamics simulations (eBDIMS) for generating accurate trajectories connecting apo and holo states³⁹. Although these coarse-grained methods predict the transition trajectory in a computationally inexpensive

manner, the knowledge of both apo and holo high-resolution structures is required a priori.

We developed a computational protocol to efficiently and accurately predict the apo $\rightarrow$ holo transition pathway, i.e., both the intermediates and holo state, solely from the apo state of a protein. This protocol is based upon linear response theory which predicts the response of a statistical mechanical system due to weak perturbations in its Hamiltonian³². Using LRT, the protocol predicts the near-equilibrium response of protein structure due to the perturbation caused by ligand binding and identifies the dominant normal modes triggered by it. After predicting the direction of conformational change, it performs an enhanced conformational search along the predicted direction to identify metastable intermediates. As protein transition pathways are typically non-linear, this protocol performs an iterative prediction of response and subsequent structure transitions to reach the end state. Although, ligand binding is a non-weak perturbation to the Hamiltonian of protein, we have also derived that linear response theory will still be valid and accurately predict the protein structural response. We have successfully tested the applicability of our protocol to three protein-ligand systems, namely, adenylate Kinase (ADK) – di(adenosine-5')pentaphosphate, ribose binding protein (RBP) – β -D-ribose, and DNA β -glucosyltransferase (BGT) – uridine-5'-diphosphate. These three systems incorporate variations in scale and nature of structural transitions upon ligand binding. For example, in ADK, a sequential hinge-bending closure of its two domains is observed whereas, in RBP, a large-scale simultaneous twisting and hinge-bending motion of two of its domains is observed. Also, the RMSD between apo and holo states varies from 2.1 Å (for BGT) to 7.1 Å (for ADK).

Theory

For a statistical mechanical system in equilibrium, LRT predicts its first order response to perturbation in its Hamiltonian⁴⁰. For a protein-ligand system, ligand binding perturbs the Hamiltonian of unbound-protein and protein change its state to reach the new free energy minima. The Hamiltonian of ligand-bound state ($\mathcal{H}_1$) can be written in terms of the Hamiltonian of ligand-unbound state ($\mathcal{H}_0$) as

$$\mathcal{H}_1 = \mathcal{H}_0 + \sum_{i=1}^N V_i \approx \mathcal{H}_0 - \sum_{j=1}^{3N} F_j X_j \quad (1)$$

V_i is the interaction energy between ligand and i^{th} residue of protein, N is the total number of residues, X_j s are the generalized coordinates such that $\{X_1, X_2, \dots, X_{3N}\} = \{x_1, y_1, \dots, z_N\}$ and F_j s are the forces exerted by the ligand on residues such that $\{F_1, F_2, \dots, F_{3N}\} = \{f_{x_1}, f_{y_1}, \dots, f_{z_N}\}$, where $\{x_i, y_i, z_i\}$ are coordinates of i^{th} residue and f_{x_i} , f_{y_i} and f_{z_i} are the ligand-exerted forces on i^{th} residue in x, y and z direction. The first order response of protein structure due to ligand-induced per-

turbation given by the LRT-formula (equation A11) is³²

$$\langle \Delta X_m \rangle = \langle X_m \rangle_1 - \langle X_m \rangle_0 = \beta \sum_{i=1}^{3N} \langle \delta X_m \delta X_i \rangle_0 F_i \quad (2)$$

$\langle \cdot \rangle_0$ and $\langle \cdot \rangle_1$ denotes ensemble averages in states before and after perturbation, $\beta (= 1/k_B T)$ is thermodynamic beta and $\langle \delta X_i \delta X_j \rangle_0$ is covariance between X_i and X_j . LRT-predicted structural response is function of fluctuation covariances of protein residues and the perturbative forces exerted by ligand (equation 2), and both the properties depends upon the conformational state of the protein. In other words, the LRT-predicted response changes as the protein conformational state changes. Generally, in large-scale conformational changes upon ligand binding, protein observes transitions from initial to final state via multiple intermediate states², and, in general, the response will change along the transition coordinate resulting in a non-linear transition pathway¹⁶ (discussed later). Therefore, it would be appropriate to use equation 2 only for the prediction of local response of protein motion from a given state to the adjacent state along the transition path but not for the prediction of the complete response due to ligand binding. Moreover, although a good correlation exists between the LRT-predicted direction of conformational change and actual direction conformational change, the magnitude or extent of conformational change is typically incorrectly estimated by the LRT-formula (Supplementary Table S1). Therefore, in our protocol we use LRT-formula to predict only the direction (not magnitude) of the structural response ($\mathbf{d}_{LRT}$) and consequently the intrinsic motions of protein in a given state triggered by ligand binding.

To identify the dominant normal modes triggered by ligand binding in a given conformational state (S), we first predicted the direction of conformational change ($\mathbf{d}_{LRT}$) using LRT in state S and then calculate the overlap (O_i) of LRT-predicted direction with each normal mode ($\mathbf{U}_i$) of state S . Overlap with i^{th} mode is calculated as

$$O_i = \mathbf{d}_{LRT} \cdot \mathbf{U}_i \quad (3)$$

Normal modes are mutually orthogonal and $\mathbf{d}_{LRT}$ can be represented by linear combination of normal modes as

$$\mathbf{d}_{LRT} = \sum_i O_i \mathbf{U}_i \quad (4)$$

Modes having a large overlap are the dominant modes triggered by the ligand binding. These dominant modes can possibly be used as CVs in enhanced sampling methods to describe the ligand-induced structural response of protein.

Besides, as LRT predicts the first order response which is accurate for weak perturbation in Hamiltonian, the question arises on the accuracy of LRT-predicted direction as ligand-binding can cause strong perturbation in protein Hamiltonian. Therefore, to examine the validity of LRT to predict direction of conformational change, we derived (*SI Index*) the general response (equation 5) of protein structure due to ligand-induced

perturbation in protein Hamiltonian which includes higher order terms.

$$\begin{aligned} \langle \Delta X_m \rangle = & \beta \sum_i \langle \delta X_m \delta X_i \rangle_0 F_i + \frac{\beta^2}{2!} \sum_{i,j} \langle \delta X_m \delta X_i \delta X_j \rangle_0 F_i F_j \\ & + \frac{\beta^3}{3!} \sum_{i,j,k} \langle \delta X_m \delta X_i \delta X_j \delta X_k \rangle_0 F_i F_j F_k + \dots \end{aligned} \quad (5)$$

Assuming a harmonic approximation of potential energy of protein, which is valid for small fluctuations near minima of an energy well, residues displacement from its mean position follows a multivariate Gaussian distribution⁴¹. For a multivariate Gaussian distribution of residue positions, we derived that (*SI Index*), each higher order term in the equation 5 is proportional to the first term and the general response is given as

$$\langle \Delta X_m \rangle = \exp(S) \times \beta \sum_{i=1}^{3N} \langle \delta X_m \delta X_i \rangle_0 F_i \quad (6)$$

$$S = \frac{\beta^2}{2} \sum_{i,j} \langle \delta X_i \delta X_j \rangle_0 F_i F_j \quad (7)$$

Equation 6 shows that the general response of protein is proportional to the linear response. Therefore, as far as the direction, and not the magnitude, of initial structural response from a given state to an adjacent state is concerned, it would be appropriate to use LRT for strong ligand-induced perturbations.

Results

Application of our protocol to predict protein structure transitions upon ligand binding is shown for three protein+ligand systems: (i) Adenylate Kinase (AdK) + Di(adenosine-5')pentaphosphate (Ap5A) (ii) Ribose Binding Protein (RBP) + β -D-ribose (RIP) and (iii) DNA β -glucosyltransferase (BGT) + Uridine-5'-Diphosphate (UDP). The main steps followed in our protocol are summarized below:

(1) Docking of ligand on protein near the binding site followed by solvation, equilibrations and 1 ns long explicit-solvent MD relaxation (see Methods).

(2) Estimation of ligand-induced force on protein residues and prediction of direction of protein structural response using LRT.

(3) Identification of dominant normal modes triggered by ligand binding and enhanced sampling along LRT-predicted direction using MDeNM.

(4) Identification of closest distinct conformational state along the excited direction and repetition of steps (2)-(4) (multiple iterations)

(5) Identification of most-stable ligand-bound state (holo state) among the output states obtained from multiple iterations

Identification of intrinsic motions triggered by ligand binding

We performed normal mode analysis (NMA) of protein fluctuations to identify the intrinsic motions of protein. NMA decompose the thermal fluctuations of protein into a set of orthogonal modes of vibration called as normal modes. We calculated normal modes by modelling protein as anisotropic network model (ANM)⁴², in which each protein residue is modelled as a single bead centred at its C α atom and each bead is connected to other beads within the cutoff radius (R_c) by harmonic springs of uniform spring constant (γ). The potential energy function (E) of this coarse-grained representation of protein is

$$E = \sum_{\forall i,j | r_{ij}^0 < R_c} \frac{\gamma}{2} (r_{ij} - r_{ij}^0)^2 \quad (8)$$

r_{ij} and r_{ij}^0 are the instantaneous and equilibrium distance between residues i and j . By carefully selecting the two parameters (R_c, γ), this elastic network model, accurately and efficiently predict slow and collective modes of protein fluctuations⁴². Relative fluctuations of residues depends upon the value of R_c , while γ only changes the overall scale of fluctuations. Thus, most important ANM parameter is R_c , which we chose on the basis of two factors:

(i) ANM should accurately predict the thermal fluctuations of protein

For this, we calculated the Pearson correlation coefficient (ρ) between residue fluctuations obtained from MD simulations and estimated using ANM for different values of R_c (see Methods). The maximum value of ρ obtained for AdK, RBP and BGT is 0.91, 0.80 and 0.78 respectively (Fig. 1, Supplementary Fig. S1, S2). High value of ρ for our test cases implies good prediction of thermal fluctuations by ANM, thus allows its use for normal mode calculations. For a given protein, we selected a set comprising all R_c values which gives correlation greater than 95 percent of the maximum correlation. The selected set of R_c values for AdK, RBP and BGT are [8 – 13 Å], [8 – 10 Å] and [8 – 10, 12 Å] respectively (Fig. 1, Supplementary Fig. S1, S2). The selected set contains R_c values of ANM which accurately predicts thermal fluctuations of protein and from this set, a single value is selected on the basis of second factor.

(ii) the identified set of relevant CVs describing the predicted direction of conformational change should be small

For each values of R_c from the selected set, we obtained the normal modes and their overlap with the LRT-predicted direction ($\mathbf{d}_{LRT}$). As discussed in theory section, modes having large overlap are the dominant modes triggered by ligand binding. These dominant modes could be selected as CVs to perform enhanced sampling and it is desirable for optimal performance of enhanced sampling methods that the set of relevant CVs should be small. In other words, since $\mathbf{d}_{LRT}$ can be represented by linear combination of normal modes (equation 4), it is desirable that most of the contribution to $\mathbf{d}_{LRT}$ is distributed among a small set of normal modes. We

Figure 1. Estimation of ANM parameters and identification of normal modes triggered by ligand binding (Adenylate Kinase). (A) Pearson correlation (ρ) between residue mean squared fluctuations predicted by ANM and estimated by 1 ns long MD simulations at different R_c values (filled circles, left Y-axis). A cutoff line at $\rho = 0.95 \times \rho_{max}$ is shown (dashed line). Mode contribution diversity estimated by η for different R_c values (empty circles, right Y-axis). (B) Mean squared fluctuations of residues in apo-state obtained from 1ns MD simulation (dotted line) and estimated by ANM at $R_c = 10$ Å (solid line). (C) Overlap (absolute value) of first 20 normal modes with LRT-predicted direction. (D) Cumulative contribution of top normal modes to LRT-predicted direction in decreasing order of overlap (circles) and reference lines at $K = 0.9, 0.95$ and 0.99 (dotted lines). (E) Direction of motion of residues in Mode 1, Mode 2 and Mode 3 shown by blue arrows on cartoon diagram of protein.

defined a Shannon-entropy based measure η to quantify the diversity of contribution of first M normal modes to $\mathbf{d}_{LRT}$.

$$\eta = \frac{S}{S_{max}} = \frac{-\sum_{i=1}^N p_i \log(p_i)}{-\sum_{i=1}^N \frac{1}{N} \log(\frac{1}{N})} \quad (9)$$

Here, $p_i = O_i^2$ as $\sum_i^M O_i^2 \approx 1$. The value of η lies between 0 and 1 for extreme cases. The case $\eta = 1$ means all M modes contribute equally to $\mathbf{d}_{LRT}$ and case $\eta = 0$ means $\mathbf{d}_{LRT}$ is completely represented by a single mode. Thus, from the selected set of R_c values, we select a single R_c which corresponds to

minimum value of η (Fig. 1, Supplementary Fig. S1, S2). Lower value of η implies most of the contribution to $\mathbf{d}_{LRT}$ comes from a small set of modes. Selected R_c value on the basis of these two factors ensures that ANM accurately estimates the intrinsic fluctuations of protein residues as well as structural response ($\mathbf{d}_{LRT}$) can be described by a small set of collective modes.

Normal modes obtained from final selected R_c value are used for further analysis. Overlap of first 20 modes with the LRT-predicted initial response is shown for AdK, RBP and BGT (Fig. 1, Supplementary Fig. S1, S2). For all test case, few slowest frequency modes have the high overlap (relevant modes) with LRT-predicted direction, whereas rest of the modes contribute very low (non-relevant modes). In order to select only relevant modes, we defined a parameter (K) to measure the cumulative contribution of top M normal modes. Value of K is zero for $M = 0$ and approaches 1 for all normal modes.

$$K = 1 - |\mathbf{d}_{LRT} - \sum_i^M O_i \mathbf{U}_i|^2 \quad (10)$$

Summation in equation 10 is over most dominant M modes in decreasing order of overlap. Smallest set of top modes whose cumulative contribution is greater than 95% is designated as the set of relevant modes triggered by ligand binding (Fig. 1(D)). We tested 3 cutoff values 90%, 95% and 99% and observed the best structure match of predicted and experimental holo-structure with 95% cutoff (data not shown). A lower cut-off value (90 %) clearly exclude some of the relevant modes, while a higher cutoff (99 %) value includes many low-overlap modes which adds noise to the calculations. Initially, 3, 3 and 2 normal modes are triggered by ligand binding in AdK, RBP and BGT respectively (Fig. 1, Supplementary Fig. S1, S2). These relevant modes are the intrinsic motions triggered by ligand binding which can potentially be used as CVs in enhanced sampling methods simulating ligand-induced protein structure transitions. We use the selected relevant modes to perform sampling along the LRT-predicted direction using MDeNM, a recently developed computationally efficient enhanced simulation method.

Enhanced sampling along LRT-predicted direction of conformational change using MDeNM

Enhanced sampling along the LRT-predicted direction of conformational change is performed using MDeNM, which is modified to perform a targeted sampling along the LRT-predicted direction. In typical MDeNM simulations, extra velocity ($\lambda \mathbf{v}_{extra}$) is added to the current velocity ($\mathbf{v}_0$) to excite protein structure along selected modes²⁴.

$$\mathbf{v} = \mathbf{v}_0 + \lambda \mathbf{v}_{extra} \quad (11)$$

Here, $\mathbf{v}$ is the updated velocity of protein, $\mathbf{v}_{extra}$ is linear combination of selected modes ($\mathbf{v}_{extra} = \sum_i \alpha_i \mathbf{U}_i$; α_i is velocity coefficient of i^{th} mode) and λ controls the degree of excitation

such that the temperature increase of system (ΔT) due to velocity excitation is

$$\Delta T = \frac{(\lambda |\mathbf{v}_{extra}|)^2}{3Nk_B} \quad (12)$$

k_B is Boltzmann constant and N is number of atoms in the system²⁴. Costa et. al. discussed the efficacy of MDeNM for random sampling ($\mathbf{v}_{extra}$ is random linear combination of selected modes) and targeted sampling along a single mode²⁴.

Generally, the LRT-predicted direction is represented by linear combination of multiple normal modes. Thus, an effective strategy is required to perform a targeted sampling along LRT-predicted using MDeNM. To perform, a targeted sampling along a pre-determined direction, velocity coefficients of normal modes need to be carefully decided. Taking velocity coefficients proportional to the predicted overlap (i.e. $\alpha_i \propto O_i$) doesn't sample the conformations in targeted direction because of unequal stiffness of modes. For example, a unit displacement along a low-frequency mode requires less excitation velocity compared to a high-frequency mode or, a given excitation velocity along the slow mode will produce a larger conformational change compared to the excitation with same velocity along a fast mode (Supplementary Fig. S3). Assuming harmonic fluctuations along a given mode i , the maximum displacement along the mode ($d_{max,i}$) is governed by the energy balance equation as

$$\frac{1}{2} v_{mean,i}^2 = \frac{1}{2} \omega_i^2 d_{max,i}^2 \quad (13)$$

ω_i is frequency of i^{th} mode and $v_{mean,i}$ is velocity at equilibrium position along i^{th} mode. Thus, based on the harmonic approximation of protein fluctuations along the normal modes, near an energy minima, we decided velocity coefficient of i^{th} mode to be proportional to the predicted overlap along the i^{th} mode (O_i) and mode frequency ω_i (i.e. $\alpha_i \propto \omega_i O_i$).

MDeNM kinetically excite the protein motion along a pre-determined direction. This excitation is equivalent to impulse, in which protein velocities are suddenly increased along the excited direction thus increasing the temperature of the system which is quickly reduced to set temperature by thermostat in simulation (Supplementary Fig. S3). Although, the extra kinetic energy is quickly dissipated by thermostat, the kinetic energy along the excited direction remains high relative to other degrees of freedom (DOF) because thermostat extract extra kinetic energy from all DOF of the system. Thus, the protein displacement along the predicted direction continues until the kinetic energy gets partitioned equally (on an average) among all DOF of the system. Therefore, in a MDeNM excitation, the initial response of protein structure is governed by the excitation temperature (ΔT) which is quickly dissipated by thermostat, subsequent response is governed by the protein's Hamiltonian or the underlying free energy landscape and total structure response is governed by both ΔT and protein's Hamiltonian. Costa et. al. discussed in detail, how protein structure changes during a MDeNM excitation.²⁴

Briefly, upon excitation, D_P increases suddenly, governed by the value ΔT , reaching a maximum value D_P^{max} within ~ 5 ps, followed by relaxation of structure (and D_P) towards the nearest energy minimum state (Supplementary Fig. S3).²⁴ More will be discussed on the characteristics of initial and total structure response with varying ΔT in the next section.

Identification and validation of protein holo-state and transition intermediates

RMSD between the C_α atoms of apo and holo crystal structures of AdK, RBP and BGT is 7.1, 6.2 and 2.1 Å respectively. After 1ns MD relaxation of docked structure (state I_1), large structure deviation remains. RMSD between C_α atoms of I_1 and holo crystal structures of AdK, RBP and BGT is 6.1, 5.2 and 3.7 Å respectively. During this small MD simulation, protein-ligand complex relax to the nearest free energy minima. Therefore, for a given protein, large perturbation in the protein structure is required to reach near-holo states, which are usually separated by large free energy barriers. Large perturbation in initial structure can be carried out using two approaches (i) a single large perturbation along the LRT-predicted direction or (ii) multiple small perturbations along the LRT-predicted directions. Importantly, the second approach allows updated calculation of direction of conformational change before each stage of small perturbation.

First approach performs a single large perturbation of initial structure and, in MDeNM, the extent of protein structure change along the predicted direction (measured as D_P = root mean square displacement of residues along excited direction; see Methods) is controlled by changing the value of excitation temperature (ΔT). However, as discussed in the theory section, D_P cannot be accurately estimated using LRT and thus, deciding the value of ΔT in advance is not possible. Therefore, we observed the protein structure change at varying degree of excitation, ranging ΔT from 0 K to 200 K at the gap of 5 K. Since, the initial response of protein in a MDeNM excitation is governed by ΔT , we observed a continuous increase in D_P^{max} with increasing ΔT . Whereas, the final response (after MD relaxation) does not increase continuously with ΔT and a three-point average shows the presence of several plateaus in D_P vs ΔT plot (Fig. 2, Supplementary Fig. S4). Plateaus corresponds to an approximately same D_P for a range of ΔT , which indicates the presence of stable region along the transition coordinate $\mathbf{d}_{LRT}$. Structures obtained in the range of ΔT between two consecutive plateaus indicates the unstable region along the transition coordinate D_P . Structures obtained in these unstable regions probably corresponds to states trapped in the local energy minimas in rugged energy landscape of the protein. Output structures obtained at the end of each excitation are compared with the experimental holo structure (Fig. 2, Supplementary Fig. S4) by calculating $RMSD$ and GDT_{avg} score (see Methods). Protein structure improvement, estimated as decrease in RMSD with the experimental holo structure after excitation, has strong correlation (0.91, 0.84 and 0.96 for ADK, RBP and BGT) with D_P till

Figure 2. Enhanced sampling along LRT-predicted direction (Adenylate Kinase). Structure displacement along LRT-predicted direction (D_P) in 100 ps MDeNM excitations at varying ΔT (circles, solid line) along with three-point average of D_P vs. ΔT data (dotted line). Identified plateau regions in average D_P vs. ΔT plot are highlighted (bottom panel). $RMSD$ (middle panel) and GDT_{avg} (top panel) score between the experimental holo structure and structure obtained at the end of MDeNM excitation. Representative structure of each plateau is highlighted (all panels).

the end of first plateau. The correlation starts to decrease for higher ΔT values and even becomes negative for very high ΔT values (Supplementary Table S2). Since protein structure improvement is strongly correlated with displacement along LRT-predicted direction, we assigned representative structures ($P_1, P_2, \dots$) to every plateau corresponding to highest D_P value in each plateau region (see Methods). Our protocol aims to predict the holo state solely from the apo state, hence, it is assumed that no information of any kind is available on holo state. Holo state is the energetically most stable ligand-bound state of a protein, thus, we identify the most-stable representative structure using a energy-based scoring function ($\Delta\Delta G$) and designate it as the predicted holo state. $\Delta\Delta G$ accounts for changes in protein-ligand binding energy and protein strain energy due to protein structure changes (see Methods). Using first approach, the RMSD between predicted holo structures and crystallographic holo structures for AdK, RBP and BGT is 3.79, 3.94 and 2.36 respectively (Table 1, Supplementary Table S3, S4).

In the second approach, large-scale structure changes, in

Figure 3. Comparison of predicted and experimental holo structures. Top panel: Cartoon diagram of initial (left) and predicted (right) holo structure (green) superimposed on experimental holo structure (Cyan); Bottom panel: Mean square displacement of predicted (black filled circles) and experimental (red empty circles) holo structure from structure I_1 along first 10 normal modes of I_1 for protein (A) AdK, (B) RBP and (C) BGT

initial state, to reach near-holo state is achieved by multiple small perturbations along the LRT-predicted directions. As the first plateau in the D_P vs ΔT plot corresponds to smallest significant transition along the LRT-predicted direction, we selected the representative structure of first plateau as the output to perform the next perturbation. Protein fluctuations, fluctuation covariance matrix and ligand-induced forces depends upon protein conformational state and thus change after each perturbation. Therefore, before each perturbation the direction of conformational change is recalculated with updated covariance matrix and ligand-induced forces. Multiple such iterations (~ 10 iterations) of prediction of conformational change and subsequent small perturbation along the predicted directions are performed. Output of each iteration is analysed using the scoring function $\Delta\Delta G$ and the most stable output state is designated as the final state. This final state is the predicted holo-state of protein and all the states connecting the transition from initial to final state are designated as intermediate states. For all three proteins, the most stable state predicted by the scoring function corresponds to the state structurally most similar to the experimental holo state as measured by $RMSD$ and GDT_{avg} score (Table 1, Supplementary Table S3, S4). $RMSD$ between predicted holo structures and experimental holo structures for AdK, RBP and BGT is 2.30, 1.58 and 1.4 Å respectively (Table 1, Supplementary Table S3, S4). Unsurprisingly, the holo structure predicted using the second approach have higher similarity with the experimental holo structure for all proteins. Large scale-structure changes, typically follow a non-linear pathways, in which intermediate states play an important role in governing and modulating conformational transitions. Since, the protein conformational transition pathways can be non-linear, the single large perturbation along the LRT-predicted direction is not the best

approach, as it ignore the important role of transition intermediates. These valid concerns are eliminated in the second approach which performs small but multiple iterative structure perturbations, incorporating the role of intermediate states in modulating the direction of transition pathway.

Fig. 3 shows the structural comparison of predicted and actual holo structures of the three proteins. Further, we calculated and compared the displacement of the predicted and experimental holo structures from the initial structure along the first 10 normal modes of the initial structure (Fig. 3). Most of the contribution to the overall displacement comes from the first few low frequency modes. Correlation between predicted and actual displacement along the first 10 modes for ADK, RBP and BGT is 0.98, 0.97 and 0.99 respectively. Although, perturbation of initial structure along a single normal mode could improve the structure similarity with the experimental holo structure, above results shows that the high similarity between predicted and actual holo structure in our case is due to correct displacement of initial structure along multiple normal modes. Apart from the validation of predicted holo structure, we analysed whether the transition intermediates/pathway predicted by our method could recapitulate the known intermediates/pathway. We selected AdK and RBP for this analysis as the information on their transition pathway and intermediates is available from past studies.

Adenylate Kinase

AdK is an ubiquitous allosteric enzyme which catalyzes the phosphorylation reaction, $AMP + ATP \rightleftharpoons 2ADP$.⁴³ This reaction involves reversible transfer of phosphoryl group from ATP (Adenosine triphosphate) to AMP (Adenosine monophosphate) which is important for cellular energy homeostasis⁸. Structure of AdK comprises three domains: CORE, LID (ATP binding domain) and NMP (AMP binding domain). AdK,

Table 1. Identification of holo structure from energetics (Adenylate Kinase).

	$RMSD (Å^0)$	GDT_{avg}	$\Delta\Delta G (Kcal/mol)$
<i>Initial</i>	7.14	0.66	-
<i>I</i> ₁	6.10	0.71	0
Single large perturbation			
<i>P</i> ₁	4.53	0.74	0.68
<i>P</i> ₂	4.02	0.77	-22.46
<i>P</i>₃	3.79	0.80	-26.04
<i>P</i> ₄	4.65	0.73	-11.00
<i>P</i> ₅	4.85	0.75	-5.62
Multiple small perturbations			
<i>E</i> ₁	4.53	0.74	0.68
<i>E</i> ₂	4.01	0.77	-0.66
<i>E</i> ₃	3.78	0.78	-12.75
<i>E</i> ₄	2.72	0.85	-29.68
<i>E</i>₅	2.30	0.87	-61.49
<i>E</i> ₆	2.49	0.85	-52.37
<i>E</i> ₇	2.75	0.83	-36.30
<i>E</i> ₈	2.76	0.83	-43.42

Initial = docked protein-ligand structure

*I*₁ = structure after 1 ns MD relaxation of *Initial*

*P*_{*i*} = representative structure of *i*th plateau

*E*_{*j*} = representative of first plateau in *j*th perturbation

upon binding to ATP and AMP, undergoes an open (inactive) → closed (active) conformational transition, in which the flexible LID and NMP domains close in towards the rigid CORE domain (Fig. 4).⁴³ Di(adenosine-5')pentaphosphate (Ap5A), an inhibitor of AdK, is commonly used to simulate the combined binding effect of AMP and ATP on AdK structure.^{8,43,44} Li and co-workers computed the free energy landscape (FEL) of AP5A-bound AdK and mapped the open → closed transition pathway using bias-exchange metadynamics simulations.⁸ We projected the transition intermediates, obtained by our protocol, in sequential order on the CV space (θ_{LID} , θ_{NMP}) defined by Li and co-workers⁸. Our predicted transition pathway is consistent with their calculations, indicating, on average, the closing of LID domain first followed by the closing of NMP domain (Fig. 4). Further, we observed a state-dependent dynamic coupling between the motions of LID and NMP domains. In state *I*₁, which is obtained after 1 ns MD relaxation of initial docked state, three normal modes are dominantly triggered by the ligand binding in the order *Mode 1* > *Mode 3* > *Mode 2* (Fig. 1(E)). Protein internal motion in all three modes involves closing of LID domain towards the CORE domain, although along the slightly different directions. Motion in *Mode 1* and *Mode 2* corresponds exclusively to the LID domain closure, motion in *Mode 3* corresponds to the coupled motion of both LID and NMP domain (Fig. 1(E)). In state *I*₁, motion in *Mode 3* shows that the closure of LID domain is dynamically coupled with opening of NMP domain. Therefore, exclusive closing motion of LID domain in *Mode 1*

Figure 4. Adenylate Kinase apo → holo transition intermediates. Transition intermediates predicted by our protocol (filled black circles) along with the minimum energy path and intermediates predicted by Li and co-workers⁸ using bias exchange metadynamics. Surface diagram of predicted protein intermediates is also shown with colour scheme: Green (CORE domain), Blue (LID domain) and Red (NMP domain), Yellow (Ligand). Definition of collective variables θ_{LID} and θ_{NMP} can be found in *SI Index*.

and *Mode 2* and anti-correlated motion of LID closure with NMP closure in *Mode 3* explains why, initially, NMP closing is unfavourable and LID closing is observed. In state *I*₄, we observed that LID domain interacts with NMP domain and the enthalpic interactions between LID and NMP domain appears to favour the NMP closure as a large closing of NMP domain is observed thereafter, *I*₄ to *I*₅ (Fig. 4). In state *I*₄, two modes are dominantly triggered by ligand binding in the order *Mode 1* > *Mode 2* (Supplementary Fig. S5). Both dominant modes triggered in state *I*₄ involves the closing of NMP domain, whereas *Mode 2* corresponds exclusively to the closing of NMP domain, whereas *Mode 1* corresponds to the twisting + closing of LID domain and closing of NMP domain. Interestingly, closing of LID domain is now correlated with the closing of NMP domain (Supplementary Fig. S5). The state-dependent coupling between the motions of LID and NMP domains explains why the characteristic non-linear pathway (LID closing followed by NMP closing) is observed in AdK upon ligand binding.

Ribose Binding Protein

RBP is a bacterial periplasmic protein which mediates ribose transport and chemotaxis.⁴⁶ It undergoes a large scale open → closed allosteric structure transition upon ribose binding,

Figure 5. Ribose Binding Protein apo $\rightarrow$ holo transition intermediates.

(A) Surface diagram of predicted protein intermediates. Residues which affect ribose transport but not ribose binding are coloured blue and ligand is shown as yellow spheres. (B) RMSD comparison of predicted intermediates with crystal structures. (C) Transition intermediates predicted by our protocol (filled circles) along with the crystallographic structures (empty circles) and metastable states of RBP predicted by Ravindranathan and co-workers⁴⁵ using umbrella sampling (stars). Definition of collective variables θ and ϕ is provided in *SI Index*.

which enables its recognition by inner membrane permease responsible for ribose transport.^{45,47} Using our protocol, we also observed an open to closed transition of RBP upon ribose binding evident by decrease in protein radius of gyration by 0.82 Å. Fig. 5(A) shows the initial, final and intermediate structures predicted for RBP transition upon ribose binding. The two domains of RBP undergoes a hinge bending and twisting motion, commonly represented by order parameters hinge angle (θ) and twist angle (ϕ) (*SI Index*). Fortunately, the crystallographic structures of RBP provides an approximate view of the path of open $\rightarrow$ closed transition in the order, 1BA2_A $\rightarrow$ 1BA2_B $\rightarrow$ 1URP $\rightarrow$ 2GX6 $\rightarrow$ 2DRI.³⁹ To compare our transition pathway with the approximate pathway predicted from crystallographic structures, we calculated the RMSD of the structures we obtained along the transition path with the above crystal structures (Fig. 5(B)). The predicted intermediate states closely matches with the crystallographic proposed intermediates with the correct order of transitions. Intermediate I_1 is closest to crystal structure 1BA2_B with rmsd of 1.39 Å, intermediate I_2 is closest to crystal structure 1URP with rmsd of 1.73 Å and final structure is closest to crystal structure 1URP with rmsd of 1.58 Å.

Conformational transition in RBP is important not only for ribose uptake, but also for ribose release upon interaction with inner-membrane-permease.⁴⁶ Ribose is almost completely buried in closed state of RBP (only 20 percent of ligand surface is exposed to solvent environment) and thus it is hypothesized to undergo a partial re-opening for ribose release.⁴⁷ The fraction of ligand surface area buried in the Initial, I_1 , I_2 , I_3 and Final states is 0.47, 0.54, 0.55, 0.75 and 0.80 respectively. From the surface area analysis, it seems that state I_2 is the closest (to final state) partially-open state which could re-

lease the ligand as approximately 55 percent of ligand-surface area is buried. Ravindranathan and co-workers determined the FEL of RBP conformational transition upon ribose binding using umbrella sampling MD simulations.⁴⁵ Our predicted transition pathway is consistent with their proposed pathway, passing close to the valley regions in FEL (Fig. 5(C)). Close to the holo state, they identified a partially-open ligand-bound state ($\theta = 122$, $\phi = 95$) on the transition path, which they proposed to be the state relevant for ribose release. This state is closest to the state I_2 ($\theta = 123$, $\phi = 79$) in the CV space, among all the intermediate states predicted by our protocol.

Discussion

One underlying hypothesis, supported by large number of experimental and computational studies, that has formed the basis of method developed here is that the structural transitions of proteins often correlate with their intrinsic motions in ligand-unbound (apo) state. Therefore, ligand binding is assumed to not introduce new large-scale motions in protein, but only amplify one or more of already existing low frequency motions. Easily accessible paths or minimum energy paths are often connected with these low frequency modes because both are associated with low free energy barriers, although still much larger than thermal energy ($\sim k_B T$). Importantly, these collective intrinsic motions can be identified easily with coarse-grained models such as ENM without explicitly sampling the free energy landscape using an all-atomistic force field. In this work, we used LRT to identify this subset of intrinsic motions that are triggered by ligand binding and used MDeNM simulations that allow crossing of barriers on the transition pathway in a short simulation by adding kinetic energy along selected normal modes. We show that excitation

temperature (ΔT) is a very important parameter that allows limiting the extent of structural change in any one step and develop a protocol for automated determination of its optimal value at each step.

Using the developed method, we predicted the structural response of three proteins: ADK, RBP and BGT upon binding to their ligands Ap5A, RIP and UDP respectively. We have shown that a small (~ 1 ns) MD relaxation after docking is not sufficient to observe the large-scale ligand-induced structure transitions. This is evident from the large RMSD between the ligand-bound structures obtained after MD relaxation of docked structure and experimental ligand-bound structures, which for the case of ADK, RBP and BGT was 6.1, 5.2 and 3.7 Å respectively. Using our approach of iterative response prediction and kinetic excitations to overcome free energy barriers, we identified the most-stable ligand-bound conformations (predicted holo-structures) along with the transition intermediates. These predicted holo-structures matches closely with the experimental holo-structure as the RMSD between the two for the case of ADK, BGT and RBP was 2.30, 1.58 and 1.40 Å respectively. We defined an energy-based scoring function, which includes changes in MM/PBSA protein-ligand interaction energy and protein strain energy, to identify the most stable ligand-bound state. For all three proteins, among all predicted conformations, the most-stable ligand-bound state identified using the energy function is structurally closest to the experimental ligand-bound structure. Further, for the case of ADK and RBP, we compared the predicted transition pathway with the transition pathway obtained from bias-exchange metadynamics and umbrella sampling simulations respectively. In the respective CVs space of both cases, we observed a close agreement between the transition pathway predicted by our approach and those obtained from the low-throughput methods in the previous studies. Crystallographic structures of RBP provides an approximate view of the path of complete transition in the order, $1BA2_A \rightarrow 1BA2_B \rightarrow 1URP \rightarrow 2GX6 \rightarrow 2DRI$. Our predicted intermediate states closely matches with the crystallographic proposed intermediates with the correct order of transitions. For example, intermediate I_1 is closest to crystal structure $1BA2_B$ with rmsd of 1.39 Å, intermediate I_2 is closest to crystal structure $1URP$ with rmsd of 1.73 Å and final structure is closest to crystal structure $1URP$ with rmsd of 1.58 Å. Also, for the case of RBP, among the intermediate states, we identified an intermediate which is possibly involved in ribose release. Interestingly, this intermediate is structurally closest to the intermediate state identified using umbrella sampling in a previous study.

To reach the final state, the method performs a series of small structure perturbations along the LRT-predicted direction. We have shown that, LRT-predicted response depends upon the conformational state of the protein and changes along the transition pathway as protein jumps from one intermediate state to another. Assuming residue fluctuations in a local free energy minima follows a Gaussian distribution, we derived

that LRT-predicted initial response in the local minima is valid even for strong ligand-induced perturbation in protein Hamiltonian. For ADK, we observed a state-dependent dynamic coupling between the motions of LID and NMP domains which explains the characteristic non-linear transition pathway (LID closing followed by NMP closing) of ADK upon binding with its ligand Ap5A. In state I_1 , which is obtained after 1 ns MD relaxation of initial docked state, closure of LID domain was anti-correlated with opening of NMP domain whereas in intermediate state I_4 , closing of LID domain was correlated with the closing of NMP domain. Since, large scale protein transition pathways are typically non-linear, intermediate states plays an important role in modulating the structural response. We compared two approaches (i) a single large structure perturbation along the LRT-predicted direction, in which the structural response is predicted only initially and role of intermediate states is not accounted and (ii) series of multiple small structure perturbations along the LRT-predicted directions, in which the response is re-calculated after each perturbation. Holo-structures predicted using the first approach has shown lower similarity with the experimental structures as compared to the second approach. In first approach, the RMSD between predicted holo structures and crystallographic holo structures for AdK, RBP and BGT was 3.79, 3.94 and 2.36 respectively. Moreover, protein structure improvement, estimated as decrease in RMSD with the experimental holo structure after excitation, has strong correlation (0.91, 0.84 and 0.96 for ADK, RBP and BGT) with structure displacement along LRT-predicted direction for small displacements. The correlation start to decrease for higher displacements and even becomes negative for very high displacement values. This suggests that LRT-predicted response is most accurate to predict small-scale response and LRT-predicted structural response should be updated along the pathway. Thus, it is important to perform small scale structure perturbations crossing only one or few energy barriers in a single step, wherein after each structure perturbation, the changes in residue fluctuations and fluctuation co-variances due to change in conformational state are incorporated.

Perhaps, the most important feature of this protocol is that, it does not require any knowledge on the end state or the reaction coordinate to predict the ligand-induced protein structure transitions. Dominant normal modes triggered by ligand binding are predicted at each intermediate steps along the transition pathway and used as CVs for structure perturbation. Although the protocol doesn't provide a continuous transition pathway but a coarse information on the characteristics of transition path is revealed by the predicted intermediate states. Interestingly, the dominantly triggered normal modes predicted at each step could be used as CVs in enhanced sampling method such as Metadynamics to gain accurate information of the transition free energy landscape. Also, the intermediate states predicted by our method could be used as inputs in path-sampling methods to obtain kinetics and thermodynamics of ligand-binding.

Methods

Protein-ligand complex preparation

Unbound protein structures are obtained from PDB database (AdK: 4AKE, BGT: 1JEJ and RBP: 1BA2_A). Ligand 3D structures are taken from crystallographic holo structures of proteins in PDB database (AdK: 1AKE, BGT: 1JG6 and RBP: 2DRI) and missing hydrogen atoms are added using Avogadro.⁴⁸ Missing hydrogen atoms in apo structure are added using the *pdb2gmx* tool of GROMACS (Version 5.1.4).⁴⁹ Ligand is docked onto the protein structure binding site using software Hex (Version 8.0.0)⁵⁰ and Autodock Vina⁵¹. Coordinates of ligand centre of mass obtained after superimposing the holo structure on apo structure are taken as reference ligand site. During ligand docking, ligand and protein structure are kept rigid and ligand translation search is constrained within a cubic box of side 40 Å centered at reference ligand site. Top 100 protein-ligand (PL) complexes, based on scoring function, are solvated with TIP3P water model⁵² and energy minimized using conjugate-gradient algorithm implemented in GROMACS⁴⁹. Ligand parameters are generated using the *Ligand Reader & Modeler* module⁵³ of CHARMM-GUI⁵⁴. After energy minimization, these 100 PL complexes are re-ranked on the basis of protein-ligand MM/PBSA⁵⁵ binding energy and most-stable PL complex complex is selected. MM/PBSA binding energies are calculated using *g_mmpbsa*⁵⁶ and detailed calculation parameters are provided in *SI Appendix*. Selected complex is equilibrated with position restraints for 100 ps under isothermal-isochoric conditions (NVT ensemble) and 100 ps under isothermal-isobaric conditions (NPT ensemble). At this stage, the system is well equilibrated at desired temperature and pressure conditions ($T = 310\text{K}$ and $P = 1\text{ atm}$) and then, we performed a 1 ns long explicit-solvent NPT production simulation to observe the allow small-scale adjustments in PL complex. Solvation, energy minimization and MD simulations are performed using CHARMM36 all-atom additive forcefield⁵⁷ implemented in GROMACS (Version 5.1.4)⁴⁹ and detailed simulation parameters of each stage are provided in *SI Appendix*. Mean squared fluctuations of residue C α positions is calculated from isotropic B-factors using equation 14, which are obtained using *rmsf* module of GROMACS on a 1 ns trajectory of apo protein in solution.

$$B_i = \frac{8\pi^2}{3} \langle \delta R_i^2 \rangle \quad (14)$$

Normal modes, residue fluctuations and covariance matrix calculation

Normal modes of a protein (N residues) in a given conformation $\mathbf{X}_0$ are calculated by representing it as an anisotropic network model (ANM)⁵⁸, in which each residue is modelled as a single bead centered at the position of its C-alpha atom and each bead in the model is connected to all other beads located within a cutoff radius R_c by harmonic springs of uniform spring constant γ . Selection of ANM parameters (R_c, γ) is discussed in Results section. Normal modes are obtained by eigendecomposition of mass-weighted Hessian matrix ($\tilde{\mathbf{H}}$).⁵⁸

$$\tilde{\mathbf{H}} = \mathbf{M}^{-\frac{1}{2}} \mathbf{H} \mathbf{M}^{\frac{1}{2}} \quad (15)$$

$\mathbf{M}$ is $3N \times 3N$ diagonal mass matrix containing mass of each residue thrice, $\mathbf{H}$ is $3N \times 3N$ Hessian matrix whose elements corresponds to the partial double derivatives of the potential energy function $E(\mathbf{X})$ and, are given as

$$H_{ij} = \left(\frac{\partial^2 E}{\partial X_i \partial X_j} \right)_{\mathbf{X}_0} \quad (16)$$

For a simplified representation, the conformation of a protein (N residues) is expressed in generalized coordinates $\mathbf{X}$ ($3N$ dimensional coordinates of C-alpha atoms of residues) such that $\mathbf{X}^T = (X_1, X_2, \dots, X_{3N})^T = (x_1, y_1, z_1, x_2, \dots, z_N)^T$. The potential energy function of ANM is given in equation 8. Eigenvectors ($\mathbf{U}_i$) of $\tilde{\mathbf{H}}^{-1}$ are the direction of normal modes and their corresponding eigenvalues (λ_i) gives the fluctuation frequencies ($\omega_i = \sqrt{\lambda_i}$) along these normal modes. First six eigenvectors (zero eigenvalues) which corresponds to the rotational and translational motion of protein are ignored. Covariance matrix of protein residue fluctuations is obtained as⁵⁸

$$\mathbf{C} = k_B T \tilde{\mathbf{H}}^{-1} \approx k_B T \sum_{i=1}^{3N-6} \frac{1}{\lambda_i} \mathbf{U}_i \mathbf{U}_i^T \quad (17)$$

Diagonal elements of $\mathbf{C}$ corresponds to variance of residue fluctuations along each axis and thus mean fluctuations of residues is estimated using ANM as⁵⁸

$$\langle \delta R_i^2 \rangle = (\langle \delta x_i^2 \rangle + \langle \delta y_i^2 \rangle + \langle \delta z_i^2 \rangle) = \sum_{k=3i-2}^{3i} C_{kk} \quad (18)$$

Calculation of ligand-induced force on residues and LRT-predicted direction of conformational change

A model based on protein-ligand MM/PBSA⁵⁵ binding energy calculations is developed to estimate the perturbative force applied by the ligand on each protein residue. Energy contribution of each residue to the total binding energy is calculated using the energy decomposition scheme of *g_mmpbsa*⁵⁶ for frames of a short MD trajectory. For any given residue, its binding energy contribution varies with change in its spatial position relative to the ligand. Thus, for each residue i , a quadratic potential energy hyper-surface model is fitted which best models its variation in binding energy contribution (ΔG_i^{BE}) vs its spatial position relative to ligand (x_i, y_i, z_i).

$$\Delta G_i^{BE} = \beta_0 + \beta_x x_i + \beta_y y_i + \beta_z z_i + \beta_{xx} x_i^2 + \beta_{yy} y_i^2 + \beta_{zz} z_i^2 + \beta_{xy} x_i y_i + \beta_{yz} y_i z_i + \beta_{zx} z_i x_i \quad (19)$$

Ligand-induced perturbative force on i^{th} residue is estimated as negative gradient of at mean residue position relative to the ligand during the trajectory.

$$\mathbf{f}_i = f_{x_i} \hat{\mathbf{x}} + f_{y_i} \hat{\mathbf{y}} + f_{z_i} \hat{\mathbf{z}} = - \left(\frac{\partial}{\partial x_i} \hat{\mathbf{x}} + \frac{\partial}{\partial y_i} \hat{\mathbf{y}} + \frac{\partial}{\partial z_i} \hat{\mathbf{z}} \right) \Delta G_i^{BE} \bigg|_{(x_i^m, y_i^m, z_i^m)} \quad (20)$$

For simplification, LRT-predicted response of all residues (equation 2), is combined together in vector form, $\Delta\mathbf{X}^T = (\langle\Delta X_1\rangle, \langle\Delta X_2\rangle, \dots, \langle\Delta X_{3N}\rangle)^T$, which is calculated as

$$\begin{bmatrix} \langle\Delta X_1\rangle \\ \langle\Delta X_2\rangle \\ \vdots \\ \langle\Delta X_{3N}\rangle \end{bmatrix} = \beta \underbrace{\begin{bmatrix} \langle\delta X_1 \delta X_1\rangle_0 & \dots & \langle\delta X_1 \delta X_{3N}\rangle_0 \\ \langle\delta X_2 \delta X_1\rangle_0 & \dots & \langle\delta X_2 \delta X_{3N}\rangle_0 \\ \vdots & \vdots & \vdots \\ \langle\delta X_{3N} \delta X_1\rangle_0 & \dots & \langle\delta X_{3N} \delta X_{3N}\rangle_0 \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_{3N} \end{bmatrix}}_{\mathbf{F}} \quad (21)$$

$\mathbf{C}$ is the covariance matrix of protein-residue fluctuations (equation 18) and F_i s are the forces exerted by the ligand on residues such that $\{F_1, F_2, \dots, F_{3N}\} = \{f_{x_1}, f_{y_1}, \dots, f_{z_N}\}$. Unit vector along $\Delta\mathbf{X}$ is referred to as the LRT-predicted direction of conformational change ($\mathbf{d}_{LRT}$).

$$\mathbf{d}_{LRT} = \frac{\Delta\mathbf{X}}{|\Delta\mathbf{X}|} \quad (22)$$

Enhanced sampling along LRT-predicted direction and selection of output structures

Enhanced sampling along the LRT-predicted direction is performed using MDeNM. Protein motion is excited along the direction which is linear combination of relevant modes. Selection of mode coefficients in the linear combination is discussed in Results section. Protein structure change is observed at varying degree of excitation. Mean residue displacement (D_P) of a protein (N residues) structure $\mathbf{X}$ from the initial structure $\mathbf{X}_0$ along the excited direction ($\mathbf{d}_{LRT}$) is calculated as

$$D_P = \sqrt{\frac{|(\mathbf{X} - \mathbf{X}_0) \cdot \mathbf{d}_{LRT}|^2}{N}} \quad (23)$$

A three-point average of D_P vs. ΔT data shows the presence of several plateaus which corresponds to the range of ΔT in which D_P fluctuates around a constant value. Mathematically, plateaus corresponds to a region of low variance in average D_P , thus we calculated a 3-point local standard deviation (LSD) of average D_P vs ΔT . Local minimas in LSD vs ΔT data corresponds to the plateau regions and two adjacent plateaus are separated by a region of high variance which are identified by local maxima between the two minimas corresponding to the adjacent plateaus. Further details (with examples) are provided in *SI Appendix*. Structure corresponding to highest D_P value in the plateau region is selected as the representative structure of the plateau.

Structure similarity comparison and scoring analysis

3D structures of a protein, after superimposition, are compared using *RMSD*, which measures the root mean squared displacement of residue $C\alpha$ positions and GDT_{avg} , which measures the average GDT (global distance test) score⁵⁹ between two structure using cutoff values of 2, 4 and 8 Å. Lower the *RMSD* or higher the GDT_{avg} between two structures, higher is the structure similarity. Mean residue displacement (D_m) of a protein (N residues) structure $\mathbf{X}$ from the initial structure

$\mathbf{X}_0$ along the m^{th} mode ($\mathbf{U}_m$) of initial structure is calculated as

$$D_m = \sqrt{\frac{|(\mathbf{X} - \mathbf{X}_0) \cdot \mathbf{U}_m|^2}{N}} \quad (24)$$

We defined a scoring function ($\Delta\Delta G$) to estimate the relative stability of predicted protein-ligand complexes compared to initial complex.

$$\begin{aligned} \Delta\Delta G &= \Delta\Delta G_{bind} + \left(E_{strain} = \sum_{m=1}^M \frac{1}{2} \omega_m^2 |(\mathbf{X} - \mathbf{X}_0) \cdot \mathbf{U}_m|^2 \right) \\ &= \Delta\Delta G_{bind} + \sum_{m=1}^M \frac{1}{2} N \omega_m^2 D_m^2 \end{aligned} \quad (25)$$

$\Delta\Delta G$ accounts for changes in protein-ligand MM/PBSA binding energy ($\Delta\Delta G_{bind}$) and protein strain energy (E_{strain}) due to changes in protein conformation. E_{strain} is estimated by assuming

Implementation

Normal mode analysis, calculation of ligand-induced force on residues and LRT-predicted direction of conformational change are performed using MATLAB scripts.

References

1. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins, DOI: [10.1038/nature06522](https://doi.org/10.1038/nature06522) (2007).
2. Amaral, M. *et al.* Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nat. Commun.* **8**, DOI: [10.1038/s41467-017-02258-w](https://doi.org/10.1038/s41467-017-02258-w) (2017).
3. Tobi, D. & Bahar, I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. United States Am.* **102**, 18908–18913, DOI: [10.1073/pnas.0507603102](https://doi.org/10.1073/pnas.0507603102) (2005).
4. Dobbins, S. E., Lesk, V. I. & Sternberg, M. J. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc. Natl. Acad. Sci. United States Am.* **105**, 10390–10395, DOI: [10.1073/pnas.0802496105](https://doi.org/10.1073/pnas.0802496105) (2008).
5. Draper-Joyce, C. & Furness, S. G. B. Conformational Transitions and the Activation of Heterotrimeric G Proteins by G Protein-Coupled Receptors. *ACS Pharmacol. Transl. Sci.* **2**, 285–290, DOI: [10.1021/acspsci.9b00054](https://doi.org/10.1021/acspsci.9b00054) (2019).
6. Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. Normal Mode Analysis of Biomolecular Structures : Functional Mechanisms of Membrane Proteins. (2009).
7. Ning, W., Fei, J. & Gonzalez, R. L. The ribosome uses cooperative conformational changes to maximize and regulate the efficiency of translation. *Proc. Natl. Acad.*

- Sci. United States Am.* **111**, 12073–12078, DOI: [10.1073/pnas.1401864111](https://doi.org/10.1073/pnas.1401864111) (2014).
8. Li, D., Liu, M. S. & Ji, B. Mapping the Dynamics Landscape of Conformational Transitions in Enzyme: The Adenylate Kinase Case. *Biophys. J.* **109**, 647–660, DOI: [10.1016/j.bpj.2015.06.059](https://doi.org/10.1016/j.bpj.2015.06.059) (2015).
9. Zheng, W. & Doniach, S. A comparative study of motor-protein motions by using a simple elastic-network model. *Proc. Natl. Acad. Sci. United States Am.* **100**, 13253–13258, DOI: [10.1073/pnas.2235686100](https://doi.org/10.1073/pnas.2235686100) (2003).
10. Guo, J. & Zhou, H. X. Protein Allostery and Conformational Dynamics, DOI: [10.1021/acs.chemrev.5b00590](https://doi.org/10.1021/acs.chemrev.5b00590) (2016).
11. Orellana, L. Large-Scale Conformational Changes and Protein Function: Breaking the in silico Barrier, DOI: [10.3389/fmolb.2019.00117](https://doi.org/10.3389/fmolb.2019.00117) (2019).
12. Kokh, D. B., Wade, R. C. & Wenzel, W. Receptor flexibility in small-molecule docking calculations. *WIREs Comput. Mol. Sci.* **1**, 298–314, DOI: [10.1002/wcms.29](https://doi.org/10.1002/wcms.29) (2011).
13. Antunes, D. A., Devaurs, D. & Kavraki, L. E. Understanding the challenges of protein flexibility in drug design, DOI: [10.1517/17460441.2015.1094458](https://doi.org/10.1517/17460441.2015.1094458) (2015).
14. Elber, R. Long-timescale simulation methods, DOI: [10.1016/j.sbi.2005.02.004](https://doi.org/10.1016/j.sbi.2005.02.004) (2005).
15. Yang, Y. I., Shao, Q., Zhang, J., Yang, L. & Gao, Y. Q. Enhanced sampling in molecular dynamics, DOI: [10.1063/1.5109531](https://doi.org/10.1063/1.5109531) (2019).
16. Tamura, K. & Hayashi, S. Linear Response Path Following: A Molecular Dynamics Method To Simulate Global Conformational Changes of Protein upon Ligand Binding. *J. Chem. Theory Comput.* **11**, 2900–2917, DOI: [10.1021/acs.jctc.5b00120](https://doi.org/10.1021/acs.jctc.5b00120) (2015).
17. Tama, F. & Sanejouand, Y.-H. Conformational change of proteins arising from normal mode calculations molecular dynamics trajectories through the quasi-harmonic. Tech. Rep. 1 (2001). DOI: [10.1093/PROTEIN/14.1.1](https://doi.org/10.1093/PROTEIN/14.1.1).
18. Yang, L. W. & Bahar, I. Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. *Structure* **13**, 893–904, DOI: [10.1016/j.str.2005.03.015](https://doi.org/10.1016/j.str.2005.03.015) (2005).
19. Tama, F., Valle, M., Frankt, J. & Brooks, C. L. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl. Acad. Sci. United States Am.* **100**, 9319–9323, DOI: [10.1073/pnas.1632476100](https://doi.org/10.1073/pnas.1632476100) (2003).
20. Wang, Y., Rader, A. J., Bahar, I. & Jernigan, R. L. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.* **147**, 302–314, DOI: [10.1016/j.jsb.2004.01.005](https://doi.org/10.1016/j.jsb.2004.01.005) (2004).
21. Bahar, I. & Rader, A. J. Coarse-grained normal mode analysis in structural biology, DOI: [10.1016/j.sbi.2005.08.007](https://doi.org/10.1016/j.sbi.2005.08.007) (2005).
22. Vashisth, H., Skiniotis, G. & Brooks, C. L. Collective variable approaches for single molecule flexible fitting and enhanced sampling, DOI: [10.1021/cr4005988](https://doi.org/10.1021/cr4005988) (2014).
23. Zheng, W. & Brooks, B. R. Normal-modes-based prediction of protein conformational changes guided by distance constraints. *Biophys. J.* **88**, 3109–3117, DOI: [10.1529/biophysj.104.058453](https://doi.org/10.1529/biophysj.104.058453) (2005).
24. Costa, M. G., Batista, P. R., Bisch, P. M. & Perahia, D. Exploring Free Energy Landscapes of Large Conformational Changes: Molecular Dynamics with Excited Normal Modes. *J. Chem. Theory Comput.* **11**, 2755–2767, DOI: [10.1021/acs.jctc.5b00003](https://doi.org/10.1021/acs.jctc.5b00003) (2015).
25. Tama, F., Miyashita, O. & Brooks, C. L. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.* **147**, 315–326, DOI: [10.1016/j.jsb.2004.03.002](https://doi.org/10.1016/j.jsb.2004.03.002) (2004).
26. Wu, X., Subramaniam, S., Case, D. A., Wu, K. W. & Brooks, B. R. Targeted conformational search with map-restrained self-guided Langevin dynamics: Application to flexible fitting into electron microscopic density maps. *J. Struct. Biol.* **183**, 429–440, DOI: [10.1016/j.jsb.2013.07.006](https://doi.org/10.1016/j.jsb.2013.07.006) (2013).
27. Wu, X., Brooks, B. R. & Vanden-Eijnden, E. Self-guided Langevin dynamics via generalized Langevin equation. *J. Comput. Chem.* **37**, 595–601, DOI: [10.1002/jcc.24015](https://doi.org/10.1002/jcc.24015) (2016).
28. Lindahl, E. & Delarue, M. Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization. *Nucleic Acids Res.* **33**, 4496–4506, DOI: [10.1093/nar/gki730](https://doi.org/10.1093/nar/gki730) (2005).
29. May, A. & Zacharias, M. Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: Evaluation on kinase inhibitor cross docking. *J. Medicinal Chem.* **51**, 3499–3506, DOI: [10.1021/jm800071v](https://doi.org/10.1021/jm800071v) (2008).
30. Cavasotto, C. N., Kovacs, J. A. & Abagyan, R. A. Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.* **127**, 9632–9640, DOI: [10.1021/ja042260c](https://doi.org/10.1021/ja042260c) (2005).
31. Yang, L. W., Kitao, A., Huang, B. C. & Go, N. Ligand-induced protein responses and mechanical signal propagation described by linear response theories. *Biophys. J.* **107**, 1415–1425, DOI: [10.1016/j.bpj.2014.07.049](https://doi.org/10.1016/j.bpj.2014.07.049) (2014).
32. Ikeguchi, M., Ueno, J., Sato, M. & Kidera, A. Protein structural change upon ligand binding: Linear response theory. *Phys. Rev. Lett.* **94**, 1–4, DOI: [10.1103/PhysRevLett.94.078102](https://doi.org/10.1103/PhysRevLett.94.078102) (2005).

33. Weiss, D. R. & Levitt, M. Can Morphing Methods Predict Intermediate Structures? *J. Mol. Biol.* **385**, 665–674, DOI: [10.1016/j.jmb.2008.10.064](https://doi.org/10.1016/j.jmb.2008.10.064) (2009).
34. Echols, N., Milburn, D. & Gerstein, M. MolMovDB: Analysis and visualization of conformational change and structural flexibility, DOI: [10.1093/nar/gkg104](https://doi.org/10.1093/nar/gkg104) (2003).
35. Lindahl, E., Azuara, C., Koehl, P. & Delarue, M. NOMAD-Ref: Visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Res.* **34**, DOI: [10.1093/nar/gkl082](https://doi.org/10.1093/nar/gkl082) (2006).
36. Franklin, J., Koehl, P., Doniach, S. & Delarue, M. MinActionPath: Maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. *Nucleic Acids Res.* **35**, DOI: [10.1093/nar/gkm342](https://doi.org/10.1093/nar/gkm342) (2007).
37. Kruger, D. M., Ahmed, A. & Gohlke, H. NMSim Web Server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins. *Nucleic Acids Res.* **40**, W310–W316, DOI: [10.1093/nar/gks478](https://doi.org/10.1093/nar/gks478) (2012).
38. López-Blanco, J. R., Aliaga, J. I., Quintana-Ortí, E. S. & Chacón, P. IMODS: Internal coordinates normal mode analysis server. *Nucleic Acids Res.* **42**, 271–276, DOI: [10.1093/nar/gku339](https://doi.org/10.1093/nar/gku339) (2014).
39. Orellana, L., Yoluk, O., Carrillo, O., Orozco, M. & Lindahl, E. Prediction and validation of protein intermediate states from structurally rich ensembles and coarse-grained simulations. *Nat. Commun.* **7**, DOI: [10.1038/ncomms12575](https://doi.org/10.1038/ncomms12575) (2016).
40. Zwanzig, R. *NONEQUILIBRIUM STATISTICAL MECHANICS* (Oxford University Press, New York, 2001).
41. Yagci, O. N., Gur, M. & Erman, B. Statistical thermodynamics of residue fluctuations in native proteins. *J. Chem. Phys.* **130**, 095103, DOI: [10.1063/1.3078517](https://doi.org/10.1063/1.3078517) (2009).
42. Atilgan, A. R. *et al.* Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80**, 505–515, DOI: [10.1016/S0006-3495\(01\)76033-X](https://doi.org/10.1016/S0006-3495(01)76033-X) (2001).
43. Arora, K. & Brooks, C. L. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc. Natl. Acad. Sci. United States Am.* **104**, 18496–18501, DOI: [10.1073/pnas.0706443104](https://doi.org/10.1073/pnas.0706443104) (2007).
44. Müller, C. W. & Schulz, G. E. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 Å resolution. A model for a catalytic transition state. *J. Mol. Biol.* **224**, 159–177, DOI: [10.1016/0022-2836\(92\)90582-5](https://doi.org/10.1016/0022-2836(92)90582-5) (1992).
45. Ravindranathan, K. P., Gallicchio, E. & Levy, R. M. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *J. Mol. Biol.* **353**, 196–210, DOI: [10.1016/j.jmb.2005.08.009](https://doi.org/10.1016/j.jmb.2005.08.009) (2005).
46. Eym, Y., Park, Y. & Park, C. Genetically probing the regions of ribose-binding protein involved in permease interaction. *Mol. Microbiol.* **21**, 695–702, DOI: [10.1046/j.1365-2958.1996.261389.x](https://doi.org/10.1046/j.1365-2958.1996.261389.x) (1996).
47. Björkman, A. J. & Mowbray, S. L. Multiple open forms of ribose-binding protein trace the path of its conformational change. *J. Mol. Biol.* **279**, 651–664, DOI: [10.1006/jmbi.1998.1785](https://doi.org/10.1006/jmbi.1998.1785) (1998).
48. Hanwell, M. D. *et al.* Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics* **4**, 17, DOI: [10.1186/1758-2946-4-17](https://doi.org/10.1186/1758-2946-4-17) (2012).
49. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25, DOI: [10.1016/j.softx.2015.06.001](https://doi.org/10.1016/j.softx.2015.06.001) (2015).
50. Ritchie, D. W. & Venkatraman, V. Ultra-fast FFT protein docking on graphics processors. *Bioinformatics* **26**, 2398–2405, DOI: [10.1093/bioinformatics/btq444](https://doi.org/10.1093/bioinformatics/btq444) (2010).
51. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, NA–NA, DOI: [10.1002/jcc.21334](https://doi.org/10.1002/jcc.21334) (2009).
52. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The J. Chem. Phys.* **79**, 926–935, DOI: [10.1063/1.445869](https://doi.org/10.1063/1.445869) (1983).
53. Kim, S. *et al.* CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules. *J. Comput. Chem.* **38**, 1879–1886, DOI: [10.1002/jcc.24829](https://doi.org/10.1002/jcc.24829) (2017).
54. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865, DOI: [10.1002/jcc.20945](https://doi.org/10.1002/jcc.20945) (2008).
55. Kollman, P. A. *et al.* Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts Chem. Res.* **33**, 889–897, DOI: [10.1021/ar000033j](https://doi.org/10.1021/ar000033j) (2000).
56. Kumari, R., Kumar, R. & Lynn, A. g_mmpbsa—A GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* **54**, 1951–1962, DOI: [10.1021/ci500020m](https://doi.org/10.1021/ci500020m) (2014).
57. Best, R. B. *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **8**, 3257–3273, DOI: [10.1021/ct300400x](https://doi.org/10.1021/ct300400x) (2012).

58. Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. *Chem. Rev.* **110**, 1463–1497, DOI: [10.1021/cr900095e](https://doi.org/10.1021/cr900095e) (2010).
59. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374, DOI: [10.1093/nar/gkg571](https://doi.org/10.1093/nar/gkg571) (2003).
60. ISSERLIS, L. ON A FORMULA FOR THE PRODUCT-MOMENT COEFFICIENT OF ANY ORDER OF A NORMAL FREQUENCY DISTRIBUTION IN ANY NUMBER OF VARIABLES. *Biometrika* **12**, 134–139, DOI: [10.1093/biomet/12.1-2.134](https://doi.org/10.1093/biomet/12.1-2.134) (1918).

Supplementary information

Prediction of direction of protein conformational change due to ligand binding

Let, a protein in solution environment is our system of interest, governed by its Hamiltonian $\mathcal{H}_0$. The partition function of the protein ($Z(\beta)$) is described by equation A1.

$$Z(\beta) \propto \int d\Gamma e^{-\beta \mathcal{H}_0} \quad (\text{A1})$$

The integral is over phase space (Γ) and $\beta = (k_B T)^{-1}$ is the thermodynamic beta. Ligand binding perturbs the Hamiltonian of unbound-protein and protein change its state to reach the new free energy minima. The Hamiltonian of ligand-bound state ($\mathcal{H}_1$) is described as perturbation to the Hamiltonian of ligand-unbound state ($\mathcal{H}_0$) (equation A2), while considering the ligand as a part of the environment.

$$\mathcal{H}_1 = \mathcal{H}_0 + \sum_{i=1}^N V_i \approx \mathcal{H}_0 - \sum_{j=1}^{3N} F_j X_j \quad (\text{A2})$$

V_i is the interaction energy between ligand and i^{th} residue of protein, N is the number of protein amino-acid residues, X_j s are the generalized coordinates such that $\{X_1, X_2, \dots, X_{3N}\} = \{x_1, y_1, \dots, z_N\}$ and F_i s are the forces exerted by the ligand on residues such that $\{F_1, F_2, \dots, F_{3N}\} = \{f_{x_1}, f_{y_1}, \dots, f_{z_N}\}$, where $\{x_i, y_i, z_i\}$ are coordinates of i^{th} residue and f_{x_i} , f_{y_i} and f_{z_i} are the ligand-exerted forces on i^{th} residue in x,y and z direction. The partition function of the perturbed protein ($Z(\beta)$) is described by equation A3.

$$Z(\beta, \{F\}) \propto \int d\Gamma e^{-\beta(\mathcal{H}_0 - \sum F_j X_j)} \quad (\text{A3})$$

A large number of system properties can be derived from the natural log of partition function ($\ln(Z(\beta, \{F\}))$). Thus, we write $\ln(Z(\beta, \{F\}))$ as a Taylor expansion with perturbation forces as variables (equation A4).

$$\begin{aligned} \ln(Z(\beta, \{F\})) &= \ln(Z(\beta, \{F\})) \Big|_{\{F\}=0} + \sum_i \frac{\partial}{\partial F_i} \ln(Z(\beta, \{F\})) \Big|_{\{F\}=0} F_i \\ &+ \frac{1}{2!} \sum_{i,j} \frac{\partial^2}{\partial F_i \partial F_j} \ln(Z(\beta, \{F\})) \Big|_{\{F\}=0} F_i F_j + \dots \end{aligned} \quad (\text{A4})$$

The first order partial derivatives in equation A4 can be derived by partial differentiation of equation A3 on both sides. The first order partial derivatives gives the ensemble averages of residue coordinates (equation A5).

$$\frac{\partial}{\partial F_i} \ln(Z(\beta, \{F\})) = \beta \frac{\int d\Gamma X_i e^{-\beta(\mathcal{H}_0 - \sum F_j X_j)}}{Z(\beta, \{F\})} = \beta \langle X_i \rangle_F \quad (\text{A5})$$

$\langle \cdot \rangle_F$ represents the ensemble average of a system property in the perturbed state. Similarly, higher order derivatives in equation A4 can be derived from equation A3. For example, second order derivatives gives the co-variance of residue position coordinates (equation A6), third order derivative gives coskewness of residue position coordinates (equation A7) and in general, the n^{th} order partial derivatives can be derived as equation A8.

$$\frac{\partial^2}{\partial F_i \partial F_j} \ln(Z(\beta, \{F\})) = \beta^2 \left(\langle X_i X_j \rangle_F - \langle X_i \rangle_F \langle X_j \rangle_F \right) = \beta^2 \langle \delta X_i \delta X_j \rangle_F \quad (\text{A6})$$

$$\frac{\partial^3}{\partial F_i \partial F_j \partial F_k} \ln(Z(\beta, \{F\})) = \beta^3 \langle \delta X_i \delta X_j \delta X_k \rangle_F \quad (\text{A7})$$

$$\frac{\partial^n}{\partial F_{i_1} \partial F_{i_2} \dots \partial F_{i_n}} \ln(Z(\beta, \{F\})) = \beta^n \langle \delta X_{i_1} \delta X_{i_2} \dots \delta X_{i_n} \rangle_F \quad (\text{A8})$$

In equation A6, A7 and A8, $\delta X_i = X_i - \langle X_i \rangle$. Since, our purpose is to predict the structural response of protein upon ligand binding, equation A5 suggests that the average coordinates of residue positions in the perturbed state can be obtained by first order partial derivatives of $\ln(Z(\beta, \{F\}))$ with respect to perturbation forces. Thus, taking a partial derivative of equation A4 with respect to F_m and multiplying both sides by $\frac{1}{\beta}$ gives

$$\begin{aligned} \frac{1}{\beta} \frac{\partial}{\partial F_m} \ln(Z(\beta, \{F\})) &= \frac{1}{\beta} \frac{\partial}{\partial F_m} \ln(Z(\beta, \{F\})) \Big|_{\{F\}=0} + \frac{1}{\beta} \sum_i \frac{\partial}{\partial F_m \partial F_i} \ln(Z(\beta, \{F\})) \Big|_{\{F\}=0} F_i \\ &+ \frac{1}{\beta} \frac{1}{2!} \sum_{i,j} \frac{\partial^2}{\partial F_m \partial F_i \partial F_j} \ln(Z(\beta, \{F\})) \Big|_{\{F\}=0} F_i F_j + \dots \end{aligned} \quad (A9)$$

Using, equation A5-A8, equation A9 can be written as

$$\langle X_m \rangle_F = \langle X_m \rangle_0 + \beta \sum_i \langle \delta X_m \delta X_i \rangle_0 F_i + \frac{\beta^2}{2!} \sum_{i,j} \langle \delta X_m \delta X_i \delta X_j \rangle_0 F_i F_j + \frac{\beta^3}{3!} \sum_{i,j,k} \langle \delta X_m \delta X_i \delta X_j \delta X_k \rangle_0 F_i F_j F_k + \dots \quad (A10)$$

Equation A10 gives the ensemble average of residue position coordinates in the perturbed state in terms of the residue coordinates before perturbation, perturbation forces, residue fluctuations and fluctuation correlations before perturbation. This equation, mathematically, shows the role of intrinsic motions in governing the conformational change of protein.

Special case 1: Weak perturbation

For a weak perturbation in protein Hamiltonian ($F_i = \text{small} \forall i$), higher order terms in equation A10 can be ignored and the resulting equation is called as linear response.

$$\langle \Delta X_m \rangle = \langle X_m \rangle_F - \langle X_m \rangle_0 \approx \beta \sum_i \langle \delta X_m \delta X_i \rangle_0 F_i \quad (A11)$$

Special case 2: Strong perturbation and Gaussian fluctuations

Assuming a harmonic approximation of potential energy of protein, which is valid for small fluctuations near minima of an energy well, residues displacement from its mean position follows a multivariate Gaussian distribution i.e. the variable $\delta \mathbf{X} = [\delta X_1, \delta X_2, \dots, \delta X_{3N}]$ is normally distributed defined by mean $\mathbf{0}$ and covariance matrix $\mathbf{C}$. For a variable with zero-mean multivariate Gaussian distribution, its higher order moments $m_n = \langle \delta X_{i_1} \delta X_{i_2} \dots \delta X_{i_n} \rangle$ can be easily determined from the covariance matrix using Isserlis's theorem (also known as Wick's probability theorem).⁶⁰ Using the theorem, when n is odd, the moment (m_n) is zero and when n is even, the moment (m_n) is given as

$$m_n = \langle \delta X_{i_1} \delta X_{i_2} \dots \delta X_{i_n} \rangle = \sum_{p \in P_n} \prod_{\{i,j\} \in p} \langle \delta X_i \delta X_j \rangle_0 \quad (A12)$$

In equation A12, P_n is a set comprising all different ways of partitioning $\{i_1, i_2, \dots, i_n\}$ into pairs $\{i, j\}$. The sum is over all different ways of partitioning and the product is over all the pairs in a partition p . Using equation A12, equation A10 can be written as

$$\begin{aligned} \langle \Delta X_m \rangle &= \beta \sum_i \langle \delta X_m \delta X_i \rangle_0 F_i + \frac{\beta^2}{2!} \sum_{i,j} \langle \delta X_m \delta X_i \delta X_j \rangle_0 F_i F_j \\ &+ \frac{\beta^3}{3!} \sum_{i,j,k} \left(\langle \delta X_i \delta X_j \rangle_0 \langle \delta X_m \delta X_k \rangle_0 + \langle \delta X_j \delta X_k \rangle_0 \langle \delta X_m \delta X_i \rangle_0 + \langle \delta X_k \delta X_i \rangle_0 \langle \delta X_m \delta X_j \rangle_0 \right) F_i F_j F_k + \dots \end{aligned} \quad (A13)$$

$$\langle \Delta X_m \rangle = \left(1 + \frac{\beta^2}{2} \sum_{i,j} \langle \delta X_i \delta X_j \rangle_0 F_i F_j + \frac{1}{2!} \frac{\beta^4}{4} \left(\sum_{i,j} \langle \delta X_i \delta X_j \rangle_0 F_i F_j \right)^2 + \dots \right) \times \beta \sum_{i=1}^{3N} \langle \delta X_m \delta X_i \rangle_0 F_i \quad (A14)$$

$$\langle \Delta X_m \rangle = \exp \left(\frac{\beta^2}{2} \sum_{i,j} \langle \delta X_i \delta X_j \rangle_0 F_i F_j \right) \times \beta \sum_{i=1}^{3N} \langle \delta X_m \delta X_i \rangle_0 F_i \quad (A15)$$

Equation A15 shows that the structural response of protein is proportional to the linear response under the assumption of Gaussian fluctuations of residue positions or, in other words, direction of structural response is same as predicted by the linear response formula (equation A11). This, implies that linear response formula (equation A11) can be used to predict the direction of initial structural response of protein even for strong ligand-induced perturbations.

Supplementary Figures and Tables

Figure S1. Estimation of ANM parameters and identification of normal modes triggered by ligand binding (Ribose Binding Protein). (A) Pearson correlation (ρ) between residue mean squared fluctuations predicted by ANM and estimated by 1 ns long MD simulations at different R_c values (filled circles, left Y-axis). A cutoff line at $\rho = 0.95 \times \rho_{max}$ is shown (dashed line). Mode contribution diversity estimated by η for different R_c values (empty circles, right Y-axis). (B) Mean squared fluctuations of residues in apo-state obtained from 1ns MD simulation (dotted line) and estimated by ANM at $R_c = 10$ Å (solid line). (C) Overlap (absolute value) of first 20 normal modes with LRT-predicted direction. (D) Cumulative contribution of top normal modes to LRT-predicted direction in decreasing order of overlap (circles) and reference lines at $K = 0.9, 0.95$ and 0.99 (dotted lines). (E) Direction of motion of residues in *Mode 1*, *Mode 2* and *Mode 3* shown by blue arrows on cartoon diagram of protein-ligand complex (ligand atoms are shown as spheres).

Figure S2. Estimation of ANM parameters and identification of normal modes triggered by ligand binding

(β -glucosyltransferase). (A) Pearson correlation (ρ) between residue mean squared fluctuations predicted by ANM and estimated by 1 ns long MD simulations at different R_c values (filled circles, left Y-axis). A cutoff line at $\rho = 0.95 \times \rho_{max}$ is shown (dashed line). Mode contribution diversity estimated by η for different R_c values (empty circles, right Y-axis). (B) Mean squared fluctuations of residues in apo-state obtained from 1 ns MD simulation (dotted line) and estimated by ANM at $R_c = 8$ Å (solid line). (C) Overlap (absolute value) of first 20 normal modes with LRT-predicted direction. (D) Cumulative contribution of top normal modes to LRT-predicted direction in decreasing order of overlap (circles) and reference lines at $K = 0.9$, 0.95 and 0.99 (dotted lines). (E) Direction of motion of residues in *Mode 1* and *Mode 2* shown by blue arrows on cartoon diagram of protein-ligand complex (ligand atoms are shown as spheres).

Figure S5. Dominant normal modes triggered by ligand in the intermediate state I_4 of AdK. Structure cartoon diagram (green), ligand (yellow sticks) and residue direction of motion (blue arrows) are shown for *Mode 1* and *Mode 2* of intermediate state I_4 obtain by our protocol during apo to holo transition of AdK

Figure S6. Adenylate Kinase transition pathway in normal mode space

Table S1. Comparison of actual and LRT-predicted magnitude of structural response

Protein	AdK	RBP	BGT
Actual <i>RMSD</i> (Å)	7.1	6.2	2.1
LRT-predicted <i>RMSD</i> (Å)	17.35	4.9	5.46

Table S2. Correlation between D_P (displacement along LRT-predicted direction) and structure improvement (measured as decrease in $RMSD$ with experimental holo structure)

ΔT range	AdK	RBP	BGT
$0 - T_1^{end}$	0.91	0.84	0.96
$T_1^{end} - T_2^{end}$	0.68	0.31	0.85
$T_2^{end} - T_3^{end}$	0.26	-0.05	-0.60
$T_3^{end} - T_4^{end}$	-0.23	-0.83	-0.96
$T_4^{end} - T_5^{end}$	-0.70	NA	NA

T_i^{end} = end of i^{th} plateau
NA = Not applicable

Table S3. Identification of holo struture from energetics (Ribose Binding Protein).

	$RMSD(A^0)$	GDT_{avg}	$\Delta\Delta G(Kcal/mol)$
<i>Initial</i>	6.2	0.64	-
I_1	5.23	0.69	0
Single large perturbation			
P_1	3.24	0.78	-0.57
P_2	3.94	0.72	-0.65
P_3	4.10	0.71	+0.63
P_4	5.49	0.62	+2.00
Multiple small perturbations			
E_1	3.24	0.78	-0.57
E_2	2.09	0.88	-1.44
E_3	1.58	0.92	-3.15
E_4	1.90	0.91	-0.98
E_5	2.42	0.87	-0.61
E_6	3.01	0.80	-0.28

Table S4. Identification of holo struture from energetics (β -glucosyltransferase).

	$RMSD(A^0)$	GDT_{avg}	$\Delta\Delta G(Kcal/mol)$
<i>Initial</i>	2.1	0.90	-
I_1	3.68	0.75	0
Single large perturbation			
P_1	2.59	0.84	-3.15
P_2	2.36	0.84	-15.26
P_3	2.67	0.79	-6.72
P_4	4.14	0.69	-0.74
Multiple small perturbations			
E_1	2.59	0.84	-3.15
E_2	1.40	0.96	-36.50
E_3	1.49	0.95	-21.67
E_4	3.04	0.80	-18.9
E_5	2.94	0.84	-3.43